Volume 12 Number 17 1984 Nucleic Acids Research

Genesis of Kirsten murine sarcoma virus: sequence analysis reveals recombination points and potential leukaemogenic determinant on parental leukaemia virus genome

John D.Norton*, Jane Connor+ and Roger J.Avery[1]

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK, and
[1]Department of Veterinary Science, Veterinary Research Laboratory, Montana State University, Bozeman, MT 59717, USA

Received 29 June 1984; Accepted 10 August 1984

ABSTRACT

The genome of Kirsten murine sarcoma virus was formed by recombination between Kirsten murine leukaemia virus sequences, and rat sequences derived from a retrovirus-like '30S' (VL30) genetic element encompassing the Kras oncogene. Using cloned DNAs we have determined the nucleotide sequences of the long terminal repeats and adjacent regions, extending across the points of recombination on the sarcoma and leukaemia virus genomes. Our results suggest that discrete regions of homology and other cryptic sequence features, may have constituted recombinational hot-spots involved in the genesis of the Kirsten murine sarcoma virus genome. We have also compared the sequence of the Kirsten murine leukaemia virus p15 env and adjacent long terminal repeat with the corresponding regions of the AKV and Gross A murine leukaemia virus genomes. This comparison has identified a leukaemogenic determinant in the U3 domain of the long terminal repeat, possibly within a enhancer-like sequence element.

INTRODUCTION

Kirsten murine leukaemia virus (KiMLV) is a chronic leukaemia virus which induces lymphomas in rodents (1,2). These neoplasms are indistinguishable from those caused by another murine leukaemia virus, Gross passage A MLV (1). In common with all other chronic leukaemia viruses the genomes of these agents do not contain an aquired cellular oncogene (3). The mechanism of tumourigenesis by this type of oncovirus is not clearly understood (4).

Kirsten murine sarcoma virus (KiMSV) was generated during the replication of KiMLV in rats (1). In contrast to KiMLV, KiMSV is an acute transforming virus and is able to rapidly induce tumours in rodents.

Extensive analysis of the KiMSV genome has shown that it arose as a result of recombination between parental KiMLV sequences and two distinct kinds of rat sequence: the 'Kras' proto-oncogene(conferring the increased oncogenicity of the virus) and a member of a class of endogenous, retrovirus-like genetic elements termed virus-like 30S (VL30)(5,6,7). The Kras oncogene is flanked by VL30 sequences within the KiMSV genome and together the two rat derived components account for about 5.0 kbp. of the 6.5 kbp. proviral DNA

© IRL Press Limited, Oxford, England.

(6,7). A similar tripartite structure is also found in the genome of the closely related Harvey murine sarcoma virus (HaMSV)(6). KiMSV and HaMSV were generated independently on infection of rats with chronic murine leukaemia viruses (MVLs). This suggests that the VL30, ras proto-oncogene and MLV progenitors possess some special feature(s) which facilitated the recombinational events involved in the genesis of the Kirsten and Harvey sarcoma viruses.

We have previously determined the location of KiMLV derived sequences on the KiMSV proviral DNA by restriction mapping and hybridization studies (7,8). This analysis showed that the parental leukaemia virus contributed the 0.5 kbp. long terminal repeats (LTRs) and an additional short sequence contiguous with the 3' LTR, to the KiMSV genome. In addition, we have recently reported the characterization and cloning of the KiMLV genome (9). During the latter study a comparison of restriction maps revealed that the KiMLV genome possesses a 3' located region (encompassing the LTR and adjacent p15 env coding sequence) very similar to that known to be responsible for the leukaemogenic potential of the Gross MLV (9,10).

In this report we describe the results of a sequence analysis of the KiMLV and KiMSV genomes, undertaken in order to investigate the molecular bases of two distinct biological processes. Firstly, information has been obtained concerning the nature of the recombination events that occurred between rat VL30 and MLV sequences during the genesis of the KiMSV genome. Secondly, we have determined the nucleotide sequence of the KiMLV genome in the region harbouring presumptive leukaemogenic determinants, in order to identify conserved sequence elements implicated in tumourigenesis.

MATERIALS AND METHODS

KiMSV DNA that had been cloned at the unique Bam HI site on circular, *In vivo* synthesized DNA (clone KCC7 - ref. 7) was excised from the vector pAT153. The sub-genomic clones of KiMLV DNA (9) cP2 and cB3 were also excised from the vector pAT153 DNA by digestion with Pst I and Bam HI respectively. Digestion with restriction enzymes (obtained from New England Biolabs) was performed under conditions recommended by the suppliers. After preparative agarose gel electrophoresis (11) various restriction fragments were ligated into the vectors M13 mp8 or M13 mp9 (12) and cloned by transformation of competent *E. coli* JM103 cells. In some instances synthetic Bam HI linkers (Collaborative Research Laboratories) were used to facilitate insertion of blunt-ended fragments. Positive plaques (clear) were identified after plating on Xgal agar plates and single stranded template DNA was prepared as

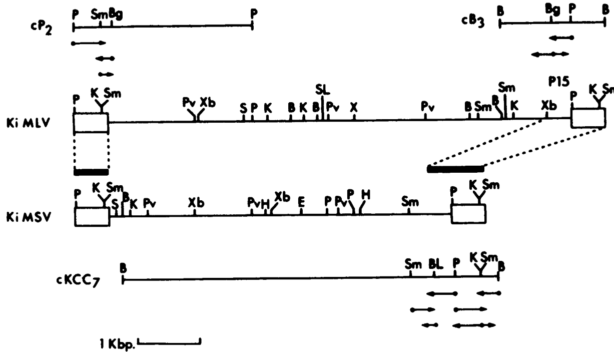

Fig. 1 Restriction maps of KiMLV and KiMSV unintegrated proviral DNAs showing the structures of recombinant clones and the nucleotide sequencing strategy. Physical maps are shown in the 5'-3' orientation with respect to viral RNA and are redrawn from refs. 7,9. The open boxes represent the LTRs and the solid black bars indicate the approximate location of KiMLV-derived sequences on the KiMSV genome. Arrows show the direction and extent of sequence analysis commencing from restriction sites indicated by dots. Restriction site abbrevations: P=Pst 1, K=KpnI, Sm=SmaI, Pv=PvuII, XbaI, S=SacI, B=Bam HI, SL=SalI, X=XhoI, E=Eco RI, H=HindIII, Bg=BglII, BL=BalI.

described elsewhere (13).

A universal 15 nucleotide primer (New England Biolabs) was used with the dideoxy chain termination method of Sanger et al. (14) to determine the sequence of M13 template inserts. Dideoxy reaction products were run on 80 cm or 40 cm polyacrylamide sequencing gels of either 4.5% or 3% monomer concentration, containing 8M urea. Gels were made up and run in standard 0.1M Tris borate buffer (pH 8.3), 2mM EDTA.

RESULTS

Fig. 1 shows the restriction maps of KiMLV and KiMSV and the approximate locations of KiMLV-derived sequences on the KiMSV genome. An approximately 1.5 kb region was sequenced on the circularly permuted clone, KCC7 (fig. 1) in order to determine the sequences of the LTR, and the KiMLV-rat VL30 recombinational boundaries. The corresponding sequence of KiMLV DNA was determined using clones P2 (LTR and 5' recombination point) and B3 (p15 env coding region and 3' recombination point).

p15 env sequence and 3' recombination point

The nucleotide sequences of the p15 env region of KiMLV and the corresponding region in KiMSV are shown in fig. 2. In the homologous region the sequences are very similar, differences are confined to a few scattered base

```
        -20       <--COOH-gp70 | NH2-p15E-->     20                 40                  60
KiMLV   AGACGAGCCAGATATAAAAAAGAACCCGTCTCACTAACTCTGGCCCTACTATTAGGAGGACTCACTATGGGCGGAATTGCC

KiMSV   GACCTTGGTGTGTAGGTTTAGATAGTGTCCAGATTAGAATCCTGATGCTAAAAAAAAATTACTTAAGGCTATCTAGGTGCT

                          80                  100                 120                 140
KiMLV   GCTGGAGTGGGAACAGGGACTACCGCCCTAGTGGCCACTCAGCAGTTCCAACAACTCCAGGCTGCCATACACGATGACCT

KiMSV   ACGAGGCAACACAAGGACATCTCCATTGCCCTACAATACAAGGTTTATAGAAAATTCCAGAACCTGCCATTGACTCAGAT

                          160                 180                 200                 220
KiMLV   TAAAGAAGTTGAAAAGTCCATCACTAATCTAGAAAAATCTTTGACCTCCTTGTCCGAAGTAGTGTTACAGAATCGTAGAG

KiMSV   ATAAAAAAGAGTGAAAGACTTTTTAGCTGACCTCCAAGGATACCTAACCTCCCTCTCAGAGGTAGTCCTTCAGAATAGGA

                          240                 260                 280                 300
KiMLV   GCCTAGATCTACTATTCCTAAAAGAGGGAGGTTTGTGTGCTGCCTTAAAAGAAGAATGCTGTTTCTATGCCGACCACACA
                                           |||| |||    |    AAGAATGCTGT|||||||||||||||||G
KiMSV   GAAGATTAGACCTGATATTCCTTAAACAAGGAACCTGTG-TGCTACACTGA

                          320                 340                 360                 380
KiMLV   GGATTGGTACGGGATAGCATGGCCAAACTTAGAGAAAGATTGAGTCAGAGACAAAAGCTCTTTGAATCCCAACAAGGGTG
KiMSV   |||||||||||A||||A||||||||||||||||||||||||A|||||||||||||||||||||A|C||||||||||||||

                          400                 420                 440                 460
KiMLV   GTTTGAAGGGCTGTTTAATAAGTCCCCTTGGTTCACCACCCTGATATCCACCGTCATGGGTCCCCTGATAATCCTCTTGT
KiMSV   ||||A|||||||||||||||||||||||||||||||||||||||||||||||A|||||||||||||||||||||||||||

                          480                 500                 520                 540
KiMLV   TAATTTTACTCTTTGGGCCTTGTATTCTCAATCGCCTGGTCCAGTTTATCAAAGACAGGATTTCGGTAGTGCAGGCCCTG
KiMSV   ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

                          560                 580 END             600                 620
KiMLV   GTTCTGACTCAACAATATCATCAACTTAAGACAATAGGAGATTAAATAAAAGATTTTATTCAGTTTACAGAAAGAGGGGG
KiMSV   ||||||||||||||||||||||||||||||||||||A||||||G||||||||||||||||||||||||||||||||||||
                  I.R.
KiMLV   GAATGAAAGACCCC --> LTR
KiMSV   ||||||||||||||
```

Fig. 2 Nucleotide sequence of the KiMLV p15 env coding region compared with the corresponding region of KiMSV DNA. The gp70-p15 env boundary was identified by comparison with that in AKV MLV (15). Nucleotides are numbered starting at the first position in the p15E coding sequence. The region around the 3' recombination point is shown boxed in. The horizontal dash indicates the absence of a nucleotide at that position. Vertical dashes indicate that the nucleotide in KiMSV is identical to that shown for the same position in KiMLV. I.R.=Inverted repeat bounding the 3' LTR.

changes. This sequence conservation allows the identification of the approximate point where KiMLV and rat VL30 progenitors recombined in the generation of the KiMSV genome. The sequence around the recombination point is shown boxed in fig. 2. Essentially complete homology is evident from position 272 (amino acid position 92 in p15 env - fig. 3) to the 3' end. By contrast the two genomes are divergent in sequence upstream from position 271.

However some limited homology occurs immediately 5' of position 272 and even further upstream we note an overall similarity in sequence between the two genomes.

The nucleotide and deduced amino acid sequence of the p15 env protein of KiMLV is shown in fig. 3, compared to the corresponding regions of the AKV and Gross A MLVs (10,15). The KiMLV sequences appear to be very similar to those of the other two viruses, which is consistent with our previous observations (9). However the carboxy terminus of the KiMLV p15 env is truncated by 5 amino acids and there is insertion of a short nucleotide sequence after the termination codon relative to the AKV and Gross A sequences.

LTR sequence and 5' recombination point

The sequence of the KiMLV LTR and adjacent downstream region is shown in fig. 4 along with the corresponding sequences of KiMSV, AKV MLV and Gross A MLV. As with the p15 env coding region, the KiMLV LTR is very similar to both the AKV and Gross A sequences. The AKV LTR has only 11 single nucleotide changes as compared to KiMLV, but does contain an additional copy of a 99 bp. sequence arranged as a tandem direct repeat. The Gross A LTR shows 15 nucleotide changes from the KiMLV sequence and has a 36 nucleotide insert not found in KiMLV. These close sequence similarities have allowed us to confidently identify various KiMLV and KiMSV sequence elements important in retroviral reverse transcription, integration and expression (reviewed in ref. 18). These elements are listed in table 1 and some of the more important landmarks are also depicted in fig. 4.

The LTR of KiMSV has only 5 nucleotides which differ from those in the LTR of its progenitor KiMLV. Downstream from the LTR the KiMSV and KiMLV genomes also have both a $tRNA^{Pro}$ primer binding site and a donor splice site in common. Immediately following the donor splice site however, the two sequences begin to diverge. As with the 3' recombination point a region of interrupted homology (shown boxed in fig. 4) occurs between the homologous and nonhomologous regions. In addition KiMLV and KiMSV sequences further downstream (approximately as far as position 1252) while not homologous, do shown an overall similarity to each other. The entire region of sequence on both genomes extending from the donor splice site downstream to position 1262 is noteworthy for its low A content. Furthermore there is also a high representation of TG dinucleotide tracts in these sequences. In KiMLV a 12 nucleotide long tract of TG residues occupies positions 1251-1262. This kind of sequence is found in eukaryotic genomes (19,20) where it is highly repeated and often flanked by the pentanucleotide TGTCT (20,21). We note that this

```
                ←—COOH-gp70 | NH2-p15E—→
        ARG ARG ALA ARG TYR LYS LYS|GLU PRO VAL SER LEU THR LEU ALA LEU LEU LEU GLY GLY    13
KiMLV   AGA CGA GCC AGA TAT AAA AAA GAA CCC GTC TCA CTA ACT CTG GCC CTA CTA TTA GGA GGA    39
                    LYS         ARG
AKV MLV  .   .   .   A   .   .   G   .   .   .   .   .   .   .   .   .   .   .   .   .

        LEU THR MET GLY GLY ILE ALA ALA GLY VAL GLY THR GLY THR THR ALA LEU VAL ALA THR    33
KiMLV   CTC ACT ATG GGC GGA ATT GCC GCT GGA GTG GGA ACA GGG ACT ACC GCC CTA GTG GCC ACT    99
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
GROSS A                                                                      .   .
MLV

        GLN GLN PHE GLN GLN LEU GLN ALA ALA ILE HIS ASP ASP LEU LYS GLU VAL GLU LYS SER    53
KiMLV   CAG CAG TTC CAA CAA CTC CAG GCT GCC ATA CAC GAT GAC CTT AAA GAA GTT GAA AAG TCC   159
                                            MET
AKV MLV  .   .   .   .   .   .   .   .   .   G   .   .   .   .   .   .   .   .   .   .
                                            MET GLN
GROSS A  .   .   .   .   .   .   .   .   .   G   G   .   .   .   .   .   .   .   .   .
MLV

        ILE THR ASN LEU GLU LYS SER LEU THR SER LEU SER GLU VAL VAL LEU GLN ASN ARG ARG    73
KiMLV   ATC ACT AAT CTA GAA AAA TCT TTG ACC TCC TTG TCC GAA GTA GTG TTA CAG AAT CGT AGA   219
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
                    LEU     ARG
GROSS A  .   .   .   T   .   G   .   .   .   .   .   .   .   .   .   .   .   .   .   .
MLV

        GLY LEU ASP LEU LEU PHE LEU LYS GLU GLY GLY LEU CYS ALA ALA LEU LYS GLU GLU CYS    93
KiMLV   GGC CTA GAT CTA CTA TTC CTA AAA GAG GGA GGT TTG TGT GCT GCC TTA AAA GAA GAA TGC   279
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
GROSS A  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
MLV

        LYS PHE TYR ALA ASP HIS THR GLY LEU VAL ARG ASP SER MET ALA LYS LEU ARG GLU ARG   113
KiMLV   TGT TTC TAT GCC GAC CAC ACA GGA TTG GTA CGG GAT AGC ATG GCC AAA CTT AGA GAA AGA   339
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
GROSS A  .   .   .   .   .   .   .   .   .   .   ARG .   .   .   .   .   .   .   .   .
MLV                                              A

        LEU SER GLN ARG GLN LYS LEU PHE GLU SER GLN GLN GLY TRP PHE GLU GLY LEU PHE ASN   133
KiMLV   TTG AGT CAG AGA CAA AAG CTC TTT GAA TCC CAA CAA GGG TGG TTT GAA GGG CTG TTT AAT   399
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
GROSS A  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
MLV

        LYS SER PRO TRP PHE THR THR LEU ILE SER THR VAL MET GLY PRO LEU ILE ILE LEU LEU   153
KiMLV   AAG TCC CCT TGG TTC ACC ACC CTG ATA TCC ACC GTC ATG GGT CCC CTG ATA ATC CTC TTG   459
                                                    ILE
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   A   .   .   .   .   .   .   .   .
                                                    ILE
GROSS A  .   .   .   .   .   .   .   .   .   .   .   A   .   .   .   .   .   .   .   .
MLV

        LEU ILE LEU LEU PHE GLY PRO CYS ILE LEU ASN ARG LEU VAL GLN PHE ILE LYS ASP ARG   173
KiMLV   TTA ATT TTA CTC TTT GGG CCT TGT ATT CTC AAT CGC CTG GTC CAG TTT ATC AAA GAC AGG   519
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .
GROSS A  .   .   .   .   .   .   .   .   .   .   .   HIS .   .   .   .   .   .   ASN .
MLV                                                  A                           A

        ILE SER VAL VAL GLN ALA LEU VAL LEU THR GLN GLN TYR HIS GLN LEU LYS THR ILE GLY   193
KiMLV   ATT TCG GTA GTG CAG GCC CTG GTT CTG ACT CAA CAA TAT CAT CAA CTT AAG ACA ATA GGA   579
                                                                                GLU
AKV MLV  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   A
                                                                                LYS
GROSS A  .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   .   AA
MLV

        ASP                     END                                    I.R.
KiMLV   GAT  -   -   -   -   -  TAAATAAAAGATTTTATTCAGTTTACAGAAAGAGGGGG[AATGAAAGACCCC]——→LTR
            CYS LYS SER ARG GLU END                                              634
AKV MLV  .  TGT AAA TCA CGT GAA ...----...............................[..........]

            CYS GLU SER ARG GLU END
GROSS A  .  TGT GAA TCA CGT GAA ...----...............................[..........]
MLV
```

same pentanucleotide sequence occurs at positions 1227-1231, and 1239-1243 in KiMLV and at positions 1227-1231, 1232-1236, and 1241-1245 in KiMSV.

Identification of the leukaemogenic determinants of KiMLV

A region of the genome encompassing the p15 env and adjacent LTR sequences has been unambiguiously identified as confering high leukaemic potential to the Gross A MLV (10). Thus a comparison of this region in the non-leukaemogenic AKV with those in the leukaemogenic Gross A and KiMLV genomes, should identify candidate determinants confering leukaemogenicity.

Within the p15 env domain, DesGroseilliers et al. have previously noted only a single position at which an amino acid is common to the leukaemogenic Gross A and Moloney MLVs, but different in the nonleukaemogenic AKV (10). This occurs at amino acid position 44 in fig. 3, where a C (nucleotide 132) in the codon for His in AKV is replaced by a G in the leukaemogenic virus, resulting in a codon for Gln. However in KiMLV the codon is the same as in AKV, implying that this position is not a leukaemogenic determinant for either Gross A or KiMLV.

Inspection of the sequences elsewhere in the p15 env coding region does not reveal any further differences from AKV which are shared by both Gross and Kirsten viruses.

Within the LTR, AKV MLV has a tandem direct repeat of 99 nucleotides. Gross A MLV differs from its presumptive parent in possessing only a single copy of this sequence, which in addition is modified by a 36 nucleotide insertion. Six point mutations further distinguish the two LTR sequences.

KiMLV does not possess the 36 bp. insertion nor does it share most of the nucleotide changes with Gross A. However at nucleotide position 684, both the Gross A and Kirsten viruses have an A residue whereas AKV MLV has a G residue. In fact this appears to be the only difference from AKV MLV, common to both Gross A MLV and KiMLV within the entire region implicated in leukaemogenicity.

DISCUSSION

The high degree of conservation of leukaemia virus derived sequences in the KiMSV genome, allowed the identification of the positions where recom-

Fig. 3 Amino acid sequence of KiMLV p15 env. The sequence is compared with AKV MLV p15 env (15) and with the coding region of Gross A MLV previously implicated as harbouring leukaemogenic determinants (10). Nucleotides and amino acids are numbered starting at the first position in the p15E coding sequence. Horizontal dashes indicate the absence of codons or nucleotides, relative to KiMLV. Vertical dashes indicate a codon or nucleotide identical to that shown for the same position in KiMLV. I.R.=Inverted repeat.

```
                I.R.                 640                 660                 680                 700
KiMLV      AATGAAAGACCCCTTCATAAGGCTTAGCAAGCTAGCTGCAGTAACGCATTTTGAAGGCATGGAAAAATACCAGAGCTGA
KiMSV      ..................................................................-............
AKV MLV    ..........................C.....A...............C....C.........G..............
GROSS A    ..........................C.....A...............C....C........................
MLV

                     720                 740                 760                 780
KiMLV      TGTTCTCAGAAAAACAAGAACAAGGAAGTACAGAGAGGCTGGAAAGTACCGGGACTAGGGCCAAACAGGATATCTGTGGT
KiMSV      ................................................................................
                                              Direct repeat
AKV MLV    ..........................[.....................................................
                                                  GAACAGATGGTCTCCAGACCACTCACTGCAGTAACG
GROSS A    ......................................AA........................................
MLV

                     800                 820                 840                 860
KiMLV      CAAACACTAGGGCCCCGGCCAGGGCCAAGAACAGATGGTCCCCAGATATAGCTAAAACAACAACAGTTTCAAGAGACCCA
KiMSV      ................................................................................
AKV MLV    ...G.............C......................A].......................................
GROSS A    T..G.............C.................C...T.....A...................................
MLV

                     880                 900                 920                 940
KiMLV      GAAACTGTCTCAAGGTTCCCCAGATGACCGGGGATCAACCCCAAGCCTCATTTAAACTAACCAATCAGCTCGCTTCTCGC
KiMSV      ................................................................................
AKV MLV    ................................................................................
GROSS A    ................................................................................
MLV

                     960                 980                 1000                1020
KiMLV      TTCTGTACCCGCGCTTATTGCTGCCCAGCTCTATAAAAAGGGTAAAAACCCCACACTCGGCGCGCCAGTCCTCCGATAGA
KiMSV      .....................................................-...........................
AKV MLV    .......................................G........................................
GROSS A    .......................................G........................................
MLV

                     1040                1060                1080                1100
KiMLV      CTGAGGCGCCCGGGTACCCGTGTATCCAATAAAGCCTTTTGCTGTTGCATCCGAATCGTGGTCTCGCTGATCCTTGGGAG
                                                           T
KiMSV      .....T......................C..............G.....................................
AKV MLV    .....T...........................................................................
GROSS A    .....T...........................................................................
MLV

                     1120            I.R. 1140   tRNA binding site 1160          1180
KiMLV      GGTCTCCTCAGAGTGATTGACTGCCCAGCTTGGGGGTCTTTCATTTGGGGGCTCGTCCGGGATTTGGAGACCCCCGCCCA
KiMSV      ................................................................................
AKV MLV    ..........................C..............END
GROSS A    ..........................C..............END
MLV

                        1200 Donor splice   1220                1240                1260
KiMLV      GGGACCACCGACCCACCGTCGGGAGGTAAGCTGGCCAGCGATCGTTTGTCTCCGTCTCTGTCTTTGTGCGTGTGTGTGTG
                             |||||||||||||| |||| |||      |||||||
KiMSV      .................TCGGGAGGTAAGCCGGCCGGCG----TTTGTCT
                                                              TGTCTGTTGTGTCTTGTCCTGTGAACGAT

                     1280                1300
KiMLV      TGCCGGGCATCTACTTTTTGCGCCTGCGTCTGAATCTGTACTAGTTAGCTAACT

KiMSV      CGATCAATAGGCTCAGATCTGGGGACTATCTGGGCGGGCCAGAGAAGGAGCTGA
```

bination occured between the rat VL30 and KiMLV progenitors of KiMSV. At both the 3' end of the KiMSV genome (within the p15 env coding domain) and near the 5' terminus (within 5', non-coding sequence) there is not a precise point at which the sequence becomes totally divergent from that of KiMLV. Rather, short regions of interrupted homology and general sequence similarity occur, before complete loss of homology. These regions of limited homology may represent the vestiges of longer sections of homology between rat VL30 and KiMLV sequences, which facilitated the recombinational events resulting in the formation of the KiMSV genome. Clearly it would be necessary to determine the sequence of the rat VL30 progenitor in order to directly assess the extent of such homology and also to identify the precise points of recombination. However, the rat VL30 gene family comprises about 150 members which exhibit considerable sequence heterogeneity (6,22). Therefore, identifying the particular VL30 genetic element which initially transduced the Kras oncogene and subsequently was itself transduced by KiMLV poses a formidable problem.

Two further sequence features of note are present near the 5' recombination point. The first of these is a common donor splice site located immediately adjacent to the point where the KiMSV and KiMLV sequences become dissimilar. The close proximity of this splice site is tantalizingly suggestive of a recombinational process that occured at the level of RNA, involving a truncated rat VL30 species onto which KiMLV 5' leader sequences were joined. In this connection it is interesting to note that we recently reported the presence of a similarly located donor splice site on a mouse VL30 genetic element (30). The presence of such a sequence on the progenitor rat VL30 element would have provided a means of generating a truncated form. However by the conventional rules of splicing such a recombinational mechanism is unlikely to have occured, since it should result in loss of the splice site. This appears to be intact, and presumably functionally essential, in the KiMSV genome.

Fig. 4 Nucleotide sequence of the KiMLV LTR and 3' adjacent region showing a comparison with KiMSV, AKV MLV (16) and Gross A MLV (17). The numbering of nucleotides is continuous with that in fig. 1. Horizontal dashes indicate the absence of nucleotides relative to KiMLV. Vertical dashes indicate a nucleotide identical to that shown for the same position in KiMLV. Nucleotide insertions relative to KiMLV are shown as superscripts at the point of insertion. The identities of various landmarks shown in the sequence are summarized in Table 1. The region around the 5' recombination point is shown boxed in. The sequence from 730 to 827 labelled direct repeat, occurs as a tandem direct repeat in the AKV genome. I.R.=Inverted repeat.

Table 1. Sequence features in the KiMLV/KiMSV LTR and adjacent regions.

FEATURE	SEQUENCE	POSITION*
+ strand primer binding site	AGAAAGAGGGGGG	609-621
5' inverted repeat	AATGAAAGACCCC	622-634
CAT box	CCAAT	921-925
Promoter	TATAAAAA	972-979
5' cap site	GC	1002
Polyadenylation signal	AATAAA	1048-1053
Polyadenylation site	CA	993
3' inverted repeat	GGGGTCTTTCATT	1133-1145
$tRNA^{Pro}$-strand primer binding site	TGGGGGCTCGTCCGGG	1146-1161
Donor splice site	AGGTAAG	1204-1210

*Positions refer to the numbering of nucleotides (beginning at the start of the p15 env coding region) adopted in figs. 2 and 4.

The second sequence feature near the 5' recombination point is the presence of extensive tracts of TG dinucleotides, flanked by copies of the sequence TGTCT. Recently, Herr (23) has reported the occurence of a very similar sequence at the same location in the genome of AKV MLV. As pointed out by this author, alternating TG dinucleotides are noteworthy in that they can exist as a left-handed Z helix under physiological conditions (24,25). This kind of sequence is highly repeated in eukaryotic genomes (19,20) and has been implicated as a hot-spot for recombination (26,27). It seems quite likely therefore that the putative homologous recombination event at the 5' side of KiMLV was facilitated by these sequence features.

Further indirect evidence that non-random, homologous recombination events were involved in the genesis of KiMSV comes from several independent studies. Firstly, Giri et al. (28) employed hybridization under relaxed conditions of stringency to demonstrate the presence of short regions of limited homology between cloned mouse and rat VL30 sequences and discrete regions of MLV genomes. These homologous regions appear to map near the inner boundaries of the MLV LTRs (28), around the positions involved in recombination with rat VL30 sequences. Secondly, in a recent sequence analysis of a retrovirus transmissible mouse VL30 genetic element (29), we directly ident-

ified limited homology with MLV genomes extending from the tRNA primer binding site up to and including the donor splice site (30). Thirdly, in genomic clones of mouse cellular DNA Itin and Keshet (31) have discovered examples of unusual VL30 genomes into which MLV sequences have been inserted. These insertions are the reciprocal of those that created the HaMSV and KiMSV genomes. Since this kind of element must have become 'fixed' in the mouse genome through germ line insertion, it follows that MLV-VL30 recombination probably occurs at a high frequency. Finally heteroduplex analysis of the Harvey and Kirsten MSV genomes failed to detect any difference in sequence at their 3' MLV-rat recombination points (5). Moreover, Goldfarb and Weinberg (32) found non-random recombination junctions between MLV and 3' truncated HaMSV sequences in the generation of biologically active HaMSV genomes. This implies that high frequency recombination occurs between specific sequences of rat VL30 and MLV genomes.

Taken together with our nucleotide sequence data, the above considerations lend support to a model of KiMLV-rat VL30 recombination in which discrete regions of homology, together with other cryptic sequence features, constituted recombinational hot-spots resulting in an ordered process of VL30 gene transduction by KiMLV.

Sequence analysis of other retroviral isolates containing an acquired cellular oncogene has, in some cases, revealed the presence of short regions of homology between the progenitor virus and cellular DNA which may have facilitated rare recombinational events (33,34). Although the precise mechanisms involved in this process are not yet established, it has been proposed that the initial event is the integration of the progenitor retroviral genome in a chromosomal location immediately upstream from the cellular oncogene. Subsequently, an RNA transcript is generated that contains both viral and cellular oncogene sequences, probably by a readthrough mechanism (particularly if the 3' LTR is absent). Packaging of this hybrid RNA species within virions of a replication competent retrovirus and infection of another cell, allows the final recombination events to occur by copy choice type mechanisms during reverse transcription (35).

We propose that in the genesis of the tripartite Kirsten and Harvey MSVs, the rat VL30 sequences supplied an adaptor function by virtue of their ability to undergo ordered high frequency recombination with MLV genomes. VL30 genetic elements are themselves transmissible by retroviruses (36) and newly acquired VL30 provirus can be detected at many integration sites following _In vitro_ infection of cells with VL30 containing MLV (A. Carter,

and J.D. Norton, unpublished). It seems probable therefore that the independent acquisition of different ras oncogenes by the rat VL30 components of HaMSV and KiMLV occured following MLV transmission of rat VL30 RNAs in infected rats. The resulting VL30 proviruses may have fortuitously integrated upstream of the respective cellular ras genes. Readthrough transcripts could then be co-packaged in MLV particles containing intact VL30 genomes. During subsequent reverse transcription events the first series or recombinations would generate VL30 units encompassing the ras oncogene. Finally, a second round of recombination could occur involving the modified VL30 elements and the "helper" MLV genomes, producing the Kirsten or Harvey MSV genomes.

By comparision of the p15 env and LTR sequences of the leukaemogenic KiMLV with the non-leukaemogenic AKV MLV we identified very few changes which could account for the difference in pathogenicity of these viruses. Only a single nucleotide substitution in the LTR at position 684 (fig. 4) was common to the leukaemogenic Gross and Kirsten viruses, but different from AKV. The possibility that the MLV genome that we have cloned and sequenced is not in fact KiMLV but another AKV-like isolate seems unlikely, since the p15 env and LTR region was found to be virtually identical to that of the derivative KiMSV genome.

Whilst this manuscript was in preparation, Lenz et al. (37) reported the localization of leukaemogenic determinants of another AKV-like MLV, SL3-3, to sequences within the LTR. This was achieved by demonstrating the leukaemogenic potential of a recombinant provirus containing the LTR of SL3-3 and the coding region of AKV MLV. These authors compared the SL3-3 LTR sequence to that of AKV and showed the determinants to map within the U3 domain (37). The SL3-3 LTR sequence appears to more closely resemble that of Gross A MLV than AKV MLV (or KiMLV). Most of the differences that were common to both Gross A and SL3-3 U3 domains occured within the direct repeat region which is known to contain enhancer elements necessary for efficient transcription (38,39). Since the activity of enhancer elements is known to be tissue specific, these authors suggested that leukaemogenicity may therefore be a function of tissue tropism. The requirement for efficient expression from the LTR promoter in the target tissue may also reflect a mechanism of tumourigenesis involving de-repression of specific cellular genes adjacent to the integrated MLV provirus (37). However within the tandem direct repeat region, the KiMLV LTR does not share with either Gross A or SL3-3 MLV any of the differences from AKV. By contrast, the G to A change at position 684 in Gross A and KiMSV is paralleled by insertion of an A nucleotide in the cor-

responding position in the SL3-3 LTR (37) suggesting that this could be of significance in determining leukaemogenicity. However, we are cautious about the interpretation of such comparative sequence data since it has not been directly demonstrated that the KiMLV LTR confers leukaemogenicity. Also this putative determinant is not within a recognized transcriptional control element as is required by the tumourigenesis model mentioned earlier. None-theless, the presence of an A nucleotide at this position is a highly conserved feature of leukaemogenic MLVs (see ref. 17). We also note that the sequence in which it occurs, TGGAAA, is very similar to the core element of an enhancer sequence (40). Conceivably, this may be a tissue specific enhancer-like element which perhaps acts synergistically with the enhancers present in the direct repeat region of U3, to promote turourigenesis by efficient expression in the target tissue.

ACKNOWLEDGEMENTS

This work was supported by grants from the Cancer Research Campaign of Great Britain. We thank Ms. V. Hill for preliminary restriction mapping of KiMSV, Ms. C. Drinville for typing and Ms. M. Mendenhall and Mrs. N. Eckenweiler for preparation of the sequence figures.

*Present address: Unit of Molecular Genetics, Liverpool School of Tropical Medicine, Pembroke Place, Liverpool L3 5QA, UK

+Present address: MRC Clinical Research Centre, Watford Road, Harrow, Middlesex HA1 3UJ, UK

REFERENCES

1. Kirsten, W.H. and Mayer, L.A. (1967) J. Nat. Cancer Inst. 39, 311-335.
2. Kirsten, W.H., Mayer, L.A., Wollmann, R.L. and Pierce, M.I. (1967) J. Natl. Cancer Inst. 38, 117-139.
3. Hardy, W.D. Nature (1984) 308, 775.
4. Teich N., Wyke J., Mak. T., Bernstein A. and Hardy W. (1982) In Molecular Biology of Tumour Viruses 2nd edn - RNA Tumour Viruses, Weiss. R., Teich N., Varmus H. and Coffin J. Eds. pp 785-998. Cold Spring Harbor Laboratory, New York.
5. Chien, Y.H., Lai, M., Shih, T.Y., Verma, I.M. Scolnick, E.M., Roy-Burman, P. and Davidson, N. (1979) J. Virol. 31, 752-760.
6. Ellis, R.W., DeFeo, D., Shih, T.Y., Gonda, M.A., Young, H.A. Tsuchida, N., Lowy, D.R. and Scolnick, E.M. (1981) Nature 292, 506-511.
7. Norton, J.D. and Avery, R.J. (1982) Biochem. Biophys. Res. Commun. 108, 1631-1637.
8. Norton, J.D., Carter, A.T. and Avery, R.J. (1982) J. Gen Virol. 58, 95-106.
9. Norton, J.D., Carter, A.T. and Avery, R.J. (1984) Biochem. Biophys. Res. Commun. 119, 150-156.

10. DesGroseillers, L., Villemur, R. and Jolicoeur, P. (1983) J. Virol. 47, 24-32.
11. Langridge, J., Langridge, P. and Bergquist, P.L. (1980) Analyt. Biochem. 103, 264-271.
12. Messing, J. and Vieira, J. (1982) Gene 19, 269-276.
13. Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
14. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Nat. Acad. Sci. (U.S.A.) 74, 5463-5467.
15. Lenz, J., Crowther, R., Straceski, A. and Haseltine, W. (1982) J. Virol. 42, 519-529.
16. Van Beveren, C., Rands, E., Chattopadhyay, S.K., Lowy, D.R. and Verma, I. M. (1982) J. Virol. 41, 542-556.
17. Villemur, R., Rassart, E., DesGroseillers, L. and Jolicoeur, P. (1983) J. Virol. 45, 539-546.
18. Chen, H.R. and Barker, W.C. (1984) Nucleic Acids Res. 12, 1767-1778.
19. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) Proc. Natl. Acad. Sci. U.S.A. 79, 6465-6469.
20. Miesfeld, R., Krystal, M. and Arnheim, N. (1981) Nucleic Acids Res. 9, 5931-5947.
21. Hamada, H., Petrino, M.G. and Kakunaga, T. (1982) Proc. Nat. Acad. Sci. U.S.A. 79, 5901-5905.
22. Young, H.A., Gonda, M.A., DeFeo, D., Ellis, R.W., Nagashima, K. and E.M. Scolnick (1980) Virol. 107, 89-99.
23. Herr, W. (1984) J. Virol. 49, 471-478.
24. Haniford, D.B. and Pulleyblank, D.E. (1983) Nature 302, 632-634.
25. Nordheim, A. and Rich, A. (1983) Proc. Nat. Acad. Sci. (U.S.A.) 80, 1821-1825.
26. Proudfoot, N.J. and Maniatis, T. (1980) Cell. 21, 537-544.
27. Stringer, J.R. (1982) Nature 296, 363-366.
28. Giri, C.P., Hodgson, C.P., Elder, P.K. Courtney, M.G. and Getz, M.J. (1983) Nucleic Acids Res. 11, 305-319.
29. Carter, A.T., Norton, J.D. and Avery, R.J. (1983) Nucleic Acids Res. 11, 6243-6254.
30. Norton, J.D., Connor, J. and Avery, R.J. (1984) Nucleic Acids Res. 12, 3445-3460.
31. Itin, A. and Keshet, E. (1983) J. Virol. 47, 178-184.
32. Goldfarb, M.P. and Weinberg, R.A. (1981) J. Virol. 38, 136-150.
33. Donoghue, D.J. and Hunter, T. (1983) J. Virol. 45, 607-617.
34. Van Beveren, C., Van Straaten, F., Curran, T., Müller, R. and Verma, I. M. (1983) Cell 32, 1241-1255.
35. Coffin, J.M. (1979) J. Gen. Virol. 42, 1-26.
36. Scolnick, E.M., Vass, W.C., Howk, R.S. and Duesberg, P.H. (1979) J. Virol. 29, 964-972.
37. Lenz, J., Celander, D., Crowther, R.L., Patarca, R., Perkins, D.W. and Haseltine, W.A. (1984) Nature 308, 467-470.
38. Laimins, L.A. Gruss, P., Pozzatti, R. and Khoury, G. (1984) J. Virol. 49, 183-189.
39. Srinivasan, A., Reddy, E.P., Dunn, C.Y. and Aaronson, S.A. (1984) Science 223, 286-289.
40. Weiher, H., Konig, M. and Gruss, P. (1983) Science 219, 626-631.